DOCUMENT RESUME

ED 462 432 TM 033 696

AUTHOR Sherman, Lawrence W.; Taylor, Amy N.
TITLE Experimentally Manipulated Bias in School Psychologists' Scoring of WISC-III Protocols.
PUB DATE 2001-10-26
NOTE 11p.; Paper presented at the Annual Meeting of the Mid-Western Educational Research Association (Chicago, IL, October 24-27, 2001).
AVAILABLE FROM For full text: http://www.users.muohio.edu/shermalw/mwera_version5_files/mwera_version5.htm.
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Bias; *Children; Down Syndrome; Elementary Secondary Education; Experiments; Intelligence Tests; Reliability; *School Psychologists; *Scoring
IDENTIFIERS *Wechsler Intelligence Scale for Children III

ABSTRACT

Experimenter bias effects were experimentally manipulated in a sample of 97 school psychologists scoring 3 subscales (Similarity, Vocabulary, and Comprehension) of the Wechsler Intelligence Scale for Children-Third Edition (WISC-III). First year (n=29), interns (n=42), and experienced (n=26) psychologists were randomly assigned to either a bias or control group and requested to score the same 3 subscale protocols. The study focused on the effects of an experimentally induced bias effect, Down Syndrome, of the subject given the WISC-III. No statistically significant interactions between experimental groups (biased and control) and level of experience were obtained. All main effects were non-significant. There results are interpreted as an affirmation of the objectivity of scoring for these relatively subjective subscales, as well as the quality of training of these students, interns, and experienced practitioners. (Contains 4 tables and 20 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made from the original document.

ED 462 432

Experimentally Manipulated Bias in School Psychologists' Scoring of WISC-III Protocols.

A presentation to the Midwest Educational Research Association
Division D: Measurement and Research Methodology
Chicago, Il
October 26, 2001

Lawrence W. Sherman (shermalw@muohio.edu) and Amy N. Taylor[1]
Department of Educational Psychology
Miami University
Oxford, Oh 45056

Available on the web at:
http://www.users.muohio.edu/shermalw/mwera_version5_files/mwera_version5.htm

TM033696

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

L Sherman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

Experimentally Manipulated Bias in School Psychologists' Scoring of WISC-III Protocols

A presentation to the Midwest Educational Research Association
Division D: Measurement and Research Methodology
Chicago, Il
October 26, 2001

Lawrence W. Sherman (shermalw@muohio.edu) and Amy N. Taylor[1]
Department of Educational Psychology
Miami University
Oxford, Oh 45056
Available on the web at:
http://www.users.muohio.edu/shermalw/mwera_version5_files/mwera_version5.htm

Abstract. Experimenter Bias Effects were experimentally manipulated in a sample of School Psychologists' (n= 97) scoring of three subscales (**Similarities, Vocabulary, Comprehension**) of the **WISC-III**. First year (***n***=29), interns (***n***=42) and experienced (***n***=26) school psychologists were randomly assigned to either a **bias** or **control** group and requested to score the identical three subscale protocols. No statistically significant interactions between experimental groups (biased vs. control) and level of experience (first-year vs. Interns vs. experienced) were obtained. All main effects were non-significant. These results were interpreted as an affirmation of the objectivity of scoring for these relatively subjective sub-scales, as well as the quality of training of these students, interns and experienced practitioners.

Intelligence tests are one integral part of educational planning and placement. The most widely used intelligence test currently on the market is the Wechsler Intelligence Scale for Children- Third Edition. Although great efforts have been made to make this test a standardized and objective measure, some subtests have been shown to be vulnerable to examiner subjectivity. Earlier research on previous versions of the WISC have indicated that several sources of bias can significantly influence an examiner's scoring of WISC-III (Sattler, 1992; Massey, 1964; Miller, 1970; Miller & Chansky, 1972; Sattler, Squire, & Andres, 1977; Slate & Chick, 1989; Slate & Jones, 1990; Slate, 1993; Wheeler, 1987; Kirchner, 1979; Shannon, 985; O'Reilly, 1989). Inasmuch as the WISC and it's subsequent revisions, the WISC-R and WISC-III, is a test that is commonly used to determine a variety of special education classifications, it would be important to know that the latest revision of this measurement device is reliable and free from bias.

Rosenthal's (1976; 1994) notion of "experimenter bias" might suggest that an examiners diagnosis of a client may unintentionally be influenced by bias, especially in the relatively subjective scoring systems associated with three specific subtests of the WISC-III. The present study focused on the effects of an experimentally induced disability bias, Down Syndrome. A randomly determined independent variable consisted of a control group not receiving this bias as contrasted with an experimental group that did receive the bias. Three levels of experience (first-year school psychology students, third-year school psychology interns, and experienced certified school psychologists) were considered as a moderator variable. Three dependent measures included the subjects' scoring of the Similarities, Vocabulary, and Comprehension subtests of the WISC-III. Based on the prior research on expectancy bias and errs observed in Wechsler scale scoring by school psychologists and trainees, the present study sought to find if a completed WISC-III protocol might also be prone to these influences. Thus, we hypothesized that our biased group would have significantly lower scores than our control group. Within the context of the many WISC sub-tests (13 in all), the three subtests that we used are often referred to as the most subjective and vulnerable to external bias. We also hypothesized that the differences between the biased and control groups would be least for the experienced school psychologists, followed by the interns and then greatest for the novice school psychology trainees.

---

[1]This study reports findings that were originally Amy N. Taylor's Specialist Degree Thesis at Miami University, Oxford, Ohio, 2001. We are thankful for Dr. Alex Thomas' support and assistance in completing this project. He greatly facilitated many of the subjects who volunteered to participate in this study.

Method

Participants:

One adolescent with Down Syndrome was selected with permission from his primary caregiver to complete a WISC-III protocol. No identifying information on this individual was known to anyone but the researchers. This individual's anonymity was protected throughout the study. Volunteers, including 29 first year school psychology graduate students, 42 intern school psychology graduate students, and 26 certified school psychologists practicing in the field were randomly assigned to either the bias or control conditions. The first year students and the interns were from various training programs in the state of Ohio including Miami University, Kent State, and the University of Akron. The certified school psychologists were randomly selected from the South Western Ohio School Psychology Association database.

Materials:

The WISC-III was given to the adolescent with Down Syndrome. The answers he gave to the subtests of Similarities, Vocabulary and Comprehension were transcribed on to a blank protocol and coded according to what group of subjects would receive the protocol (e.g. bias vs. control and level of training). This was done so that the researcher could identify which group each protocol belonged to without identifying any subjects, who for the most part remained anonymous. The protocols were coded by changing the response on one answer on the Similarities subtest slightly according to which group they belonged to. Half of the transcriptions to each group included a sheet of paper indicating that the individual who completed this protocol had Down Syndrome while the other half received no information with the protocol. These letters told the subjects nothing about the study's purpose. They were only told to score the subtests as best they could without using the WISC-III manual or any other aid. The subjects were told not to use the manual to prevent sharing of the manual by the subjects, which could lead to collaboration on responses, and thereby tainting the data collected.

Procedure:

One of the researchers first gave the WISC-III subtests of Similarities, Vocabulary and Comprehension to the adolescent with Down syndrome. When the protocol was completed, the researcher then transcribed the answers to the questions on the Similarities, Vocabulary, and Comprehension subtests onto six different blank protocols. These six groups were coded according to which group they belonged, and copies of this transcription were then made. A coding system was developed and placed on the protocol to indicate level of training of the subjects and whether or not they received the bias to facilitate a "double-blind" element of the study. Half of the subjects from each level of training randomly received the bias and the other half did not. The bias consisted of a small sheet of paper stating that an individual with Down Syndrome completed the three subtests they received. All subjects also received instructions with their transcriptions asking them to score the subtests without using any scoring guide and to not share their answers or protocols with anyone else to insure confidentiality. They were also instructed to mail their completed protocols back to the researchers in the enclosed self-addressed envelopes via the Educational Psychology office at Miami University. No identifying information was placed on any protocol to identify the subjects who remained anonymous. The researcher was only aware of whether the subject was a first year graduate student, an intern or a certified school psychologist and whether or not they received or did not receive the bias. Thus, the researchers were "blind" as to who received what experimental condition. The researchers then processed each subjects' raw scores for each subtest and used them as dependent measures in later analyses.

Research Design:

This study was a randomized posttest-only control group design (see Table 1). The subjects in the three different groups (student, intern, certified - the moderating variable) were randomly assigned to

receive the bias or not receive the bias - the independent variable. The posttest consisted of determining if there was a significant difference in raw scores (the dependent variable) among the groups receiving the bias or not receiving the bias - the independent variable. Further, the researchers sought to determine if these score differences varied depending on the three experience levels of the group (the moderating variable). We hypothesized that the mean scores for the biased groups would be significantly lower on each of the three subtests than the control groups' mean scores. Further, the difference between scores in the biased versus control conditions would be greatest for the first year students, then the interns and least for the certified school psychologists (see Table 2).

Table 1. Symbolic Representation of Research Design.

	Similarities	Vocabulary	Comprehension
Group I: Students	R_n—X—O_1	R_n—X—O_1	R_n—X—O_1
	R_n—C—O_2	R_n—C—O_2	R_n—C—O_2
Group II: Interns	R_i—X—O_3	R_i—X—O_3	R_i—X—O_3
	R_i—C—O_4	R_i—C—O_4	R_i—C—O_4
Group III: Certified	R_c—X—O_5	R_c—X—O_5	R_c—X—O_5
	R_c—C—O_6	R_c—C—O_6	R_c—C—O_6

R_n= Randomly selected first year graduate students in School Psychology (stu)

R_i= Randomly selected interns (third year of study) in School Psychology (int)

R_c= Randomly selected certified School Psychologists (cer.)

X= Bias given (treatment)

C= Control (no bias given

O= Scores

Table 2: Research Hypotheses I and II

Hypothesis I	Hypothesis II
$O_1 < O_2$	$[O_2-O_1]>[O_4-O_3]>[O_6-O_5]$
$O_3 < O_4$	
$O_5 < O_6$	

O_1= Scores for experimental group I

O_2= Scores for control group I

O_3= Scores for experimentally group II

O_4= Scores for control group II

O_5= Scores for experimental group III

O_6= Scores for control group III

Results.

The experimental treatment, bias vs. control group, did not significantly interact with the level of experience on any of the three sub-tests. These results are reported in Tables 3 and 4. No significant main effects were obtained on the Similarities or Vocabulary sub-tests. The Comprehension sub-test did obtain one marginally significant main effect for level of experience ($F_{(2, 91)} = 3.24$, $p < .02$), with first year students scoring their protocols significantly (Scheffe = 3.14) higher than the experienced practitioners who scored their protocols the lowest, and the third-year school psychology interns contributed scores in the middle, not significantly different from the first year students or experienced practitioners.

Table 3: Subscale Means, Standard deviations, and Ranges

	Subscales					
	Similarities		Vocabulary		Comprehension	
Groups	Bias	Control	Bias	Control	Bias	Control
Students (n=29)	n = 13	n=16	n=13	n=16	n=13	n=16
Mean	11.62	12.19	14.77	15.75	14.77	15.00
SD	.92	1.90	4.64	3.53	3.39	3.92
Interns (n=42)	n =21	n =21	n =21	n =21	n =21	n =21
Mean	12.38	12.48	16.71	16.24	14.29	13.29
SD	1.94	1.29	2.97	1.95	2.59	2.45
Certified (n=26)	n=12	n=14	n=12	n=14	n=12	n=14
Mean	11.75	11.57	16.17	17.14	12.00	13.57
SD	1.52	.934	2.48	1.96	2.59	3.37
ANOVA INTERACTION	$F_{(2,91)}$ =	.34, ns	$F_{(2,91)}$ =	.61, ns	$F_{(2,91)}$ =	1.44, p< .26

Table 4: Three 2x3 ANOVAs* of Experimental Treatment (Bias/Unbiased) by Experience (3) for Three WISC-III Subtests

Subscale	Source	df	MS	F	P
Similarities					
	Bias/Unbiased	1	.61	.21	.64
	Experience Level	2	5.27	1.84	.64
	Interaction	2	1.00	.34	.70
	Error	91	2.87		
Vocabulary					
	Bias/Unbiased	1	5.63	.62	.42
	Experience Level	2	17.76	1.98	.14
	Interaction	2	5.43	.61	.55
	Error	91	8.95		
Comprehension					
	Bias/Unbiased	1	1.65	.17	.67
	Experience Level	2	33.96	3.66	.02
	Interaction	2	12.74	1.37	.26
	Error	91	9.28		

*Statistics computed using GB-STAT (Friedman, a1998).

Discussion

No significant differences in scoring were found for the experimental versus the control groups on the Similarities, Vocabulary or Comprehension subtests. Bias alone showed no effect on any of the three subtests. However, Level of experience did show a marginally significant effect on the Comprehension subtest. This was not a hypothesis that was originally to be tested by the researchers, but an interesting finding nonetheless, even if somewhat marginal in significance. However, this finding did not take the direction the researchers would have expected. On the Comprehension subtest, the certified School Psychologists produced significantly lower means than the first year students. This finding demonstrates that level of experience may be related to differential scoring on this subtest. It does not support the researchers' hypothesis that the more experienced an individual is, the less likely they will be influenced by a bias. This may be attributed to experienced school psychologists being more stringent in their scoring or to novice school psychology trainees being too liberal. Technically, if one is to

assume that no matter what level of experience a professional is at, they still will have mastered the scoring techniques, then experience alone should not have had a significant effect. This finding could speak to the test-makers about searching out ways to make the scoring of this subtest less subjective and therefore examiners more capable of arriving at a uniform score.

This study was limited in several ways. The small number of subjects and the limitation in geographical representation subsequently makes the obtained results less generalizable. Further research should concentrate on broadening both the size of the groups as well as the geographic and demographic diversity of the subjects involved. The final and perhaps most obtrusive limitation of this study was the overall contrived nature of the research. Having school psychologists and school psychologists in training score protocols from a child they have never seen, much less assessed, is very unrealistic. In the real world of practice, the child would be in front of the examiner and a much "truer" score would likely be determined. On the other hand, having a child with Down Syndrome in front of the examiner may even heighten the effect of the bias given the possible influence of observed physical attributes of the child being assessed. One can never really know what the "true" effects would be.

While this study failed to confirm hypotheses based on Rosenthal's (1994) experimenter bias effect, the results are interpreted as an affirmation of the objectivity of scoring for these relatively subjective sub-scales, as well as the quality of training of these students, interns and experienced practitioners.

References

Beeghly, M. & Cicchetti, D. (1990). *Children with Down syndrome: A developmental perspective*. Cambridge University Press: Boston, MA.

Braden, J.P. (1995). Review of the WISC-III. In J.C. Conoley & J. C. Impara (Eds.), *The twelfth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements. See Test number 412.

Friedman, P. (1998). *GB-STAT Tutorial*. Silver Spring, MD: Dynamic Microsystems, Inc.

Kirchner, C. M. (1980). Children's test behavior and examiner bias on the Wechsler Intelligence Scale for Children. *Dissertation Abstracts International*-A, 41/06, p. 2566.

O'Reily, C. (1989). The confirmation bias in special education eligibility decisions. *School Psychology Review, 18*, (n 1), 126-35.

Massey, J.O. (1964) *WISC scoring criteria*. Palo Alto, CA: Consulting Psychologists Press.

Miller, C.K., Chansky, N.M. & Gredler, G.R.(1970). Rater agreement on WISC protocols. *Psychology in the Schools*, 7, 190-193.

Miller, C.K. & Chansky, N.M. (1972). Psychologists' scoring of WISC protocols. *Psychology in the Schools*, 9, 144-152.

Rosenthal, R. (1976). *Experimenter Effects in Behavioral Research*. Irvington Publishers, Inc.: New York, NY.

Rosenthal, R. (1994). Interpersonal expectancy effects: A 30-year perspective. *Current Directions in Psychological Science*, 3, 176-179.

Sandoval, J. (1995). Review of the WISC-III. In J.C. Conoley & J. C. Impara (Eds.), The twelfth mental measurements yearbook. Lincoln, NE: Buros Institute of Mental Measurements. See Test number 412.

Sattler, J.M., Squire, L.S., & Andres, J.R. (1977). Scoring discrepancies between the WISC-R manual and two scoring guides. *Journal of Clinical Psychology*, 33, 1058-1059.

Sattler, J.M. (1992). *Assessment of Children-Revised and Updated, 3rd ed.* Jerome M. Sattler Publisher: San Diego, CA.

Shannon, Robert Lewis (1985). Impact of special education labels: Implications for psychological re-evaluations utilizing the WISC-R. *Dissertation Abstracts International-A*, 45/08, p. 2488.

Shaw, S.R., Swerdlik, S.E., & Laurent, J. (1993). [Review of the WISC-III.] In B.A. Bracken (Ed.), *Monograph series advances in psychoeducational assessment: Wechsler Intelligence Scale for Children-Third edition; Journal of Psychoeducational Assessment* (pp. 151-159). Brandon, VT: Clinical Psychology Publishing Co., Inc.

Slate, J.R. (Jan 1993). Evidence that practitioners err in administering and scoring the WAIS-R. *Measurement and Evaluation in Counseling and Development*, 25,156-161.

Slate, J.R. & Chick, D. (Jan 1989). WISC-R examiner errors: Cause for concern. *Psychology in the Schools*, 26, 78-84.

Slate, J.R. & Jones, C.H. (1990a). Identifying students' errors in administering the WAIS-R. *Psychology in the Schools*, 27, 83-87.

Slate, J. R. & Jones, C.H. (1990b). Student error in administering the WISC-R: Identifying problem areas. *Measurement and Evaluation in Counseling and Development*, 23, 137-140.

Taylor, Amy N. (2001). *Experimentally Manipulated Bias in School Psychologists' Scoring of WISC-III Protocols* , (a Specialist's Degree Thesis), Miami University, Oxford, OH.

Wheeler, P. T. (1987). A study of the effect of a child's physical attractiveness upon verbal scoring of the Wechsler Intelligence Scale for Children (Revised) and upon personality attributions. *Dissertation Abstracts International-B*, 47/08, p. 3550.